Summary: This paper summarizes exploratory data analysis for work investigating student software language preferences and student interest in potential new courses in order to guide software, systems and data science curriculum planning.

Research design: Survey Monkey collects **20 useful attributes** from 207 students. Via 3 constant sum questions, students attribute personal desire, professional need, and industry importance to 5 language categories resulting in 3 x 5 = **15 metric language preference attributes**. Via 4 sliding scales, students vote their interest in 4 courses referred to as **4 metric course interest attributes**. Via drop down students indicate **number of courses completed**.

Table 1:
Number and % of column missing values in the data set:

	MSPA na	MSPA %na
My_Java	0	0
My_JS	0	0
My_Py	0	0
My_R	0	0
My_SAS	0	0
Pro_Java	0	0
Pro_JS	0	0
Pro_Py	0	0
Pro_R	0	0
Pro_SAS	0	0
Ind_Java	0	0
Ind_JS	0	0
Ind_Py	0	0
Ind_R	0	0
Ind_SAS	0	0
Py take	0	0
DE take	5	3
AA take	2	1
SA take	5	3
# taken	17	10
PREDICT400	33	20
PREDICT401	26	15
PREDICT410	46	27
PREDICT411	75	45
PREDICT413	119	72
PREDICT420	60	36
PREDICT422	125	75
PREDICT450	151	91
PREDICT451	161	97
PREDICT452	155	93
PREDICT453	155	93
PREDICT454	160	96
PREDICT455	140	84
PREDICT456	159	96
PREDICT457	161	97
otherPy	160	96
OtherR	154	93
OtherSAS	163	98
Other	144	87
Graduate_Date	2	1

Table 2: Skew, Median, Mean compared for imputation:

	skew	median	mean
My_Java	1.424146	5.0	10.387879
My_JS	1.326340	0.0	4.581818
My_Py	0.820214	30.0	31.557576
My_R	0.985261	35.0	36.933333
My_SAS	0.579966	15.0	16.539394
Pro_Java	2.155574	5.0	9.575758
Pro_JS	4.607965	0.0	5.781818
Pro_Py	0.726639	30.0	30.460606
Pro_R	0.821401	33.0	35.593939
Pro_SAS	1.251565	20.0	18.587879
Ind_Java	1.515195	5.0	11.660606
Ind_JS	1.752009	0.0	6.678788
Ind_Py	0.672794	30.0	30.151515
Ind_R	0.441026	30.0	33.212121
Ind_SAS	1.182659	15.0	18.296970
Py take	-1.098666	82.0	73.806061
DE take	-0.260672	60.0	57.890909
AA take	-0.203657	61.0	56.096970
SA take	-0.092586	54.5	54.645455
# taken	0.147758	6.0	6.333333
All_Py_Index	8.263390	0.6	1.030934

Exploratory Data Analysis: Our 41-column dataset contains a respondent id (unique row identifier), the 20 useful attributes, and 20 discarded columns that include 18 checkboxes and a single textbox that are missing 20 – 98% of their values **(see table 1 above)**. The final attribute, Graduate_Date is discarded in favor of the courses completed attribute: while the former is missing only 1% of its rows and the latter 10%, the latter is ratio level factual data whereas the former reflects student expectations that might be subjective. This discussion of the **training set's univariate** descriptors provides some methodological guidance. Firstly, while 3 constant sum survey responses are ratio level data WITHIN each **language preference question**, they cannot be compared across questions. In any case, this survey the results are roughly uniform across all 3 contexts: personal, professional, industry. See **table 3**: Python and R median spends range 30-35% or twice the median 11-16% spent each for SAS and Java. The 3 Javascript

Table 3: Univariate measures for language importance:

	My_JS	My_Py	My_R	My_SAS	Pro_Java	Pro_JS	Pro_Py	Pro_R	Pro_SAS	Ind_Java	Ind_JS	Ind_Py	Ind_R	Ind_SAS	All_Py_Index
count	165.0	165.0	165.0	165.0	165.0	165.0	165.0	165.0	165.0	165.0	165.0	165.0	165.0	165.0	165.0
miss	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
min	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
max	25.0	90.0	100.0	50.0	80.0	100.0	100.0	100.0	100.0	70.0	50.0	95.0	85.0	90.0	26.5
mean	4.6	31.6	36.9	16.5	9.6	5.8	30.5	35.6	18.6	11.7	6.7	30.2	33.2	18.3	1.0
median	0.0	30.0	35.0	15.0	5.0	0.0	30.0	33.0	20.0	5.0	0.0	30.0	30.0	15.0	0.6
std	6.6	15.5	14.7	13.6	13.8	11.0	18.8	19.8	18.7	15.1	9.7	18.4	16.1	18.3	2.5
skew	1.3	0.8	1.0	0.6	2.2	4.6	0.7	0.8	1.3	1.5	1.8	0.7	0.4	1.2	8.3
kurt	0.7	2.3	2.4	-0.2	6.1	33.6	1.4	1.6	2.3	2.0	3.6	1.1	0.8	1.4	75.1
01p	0.0	0.0	10.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
05p	0.0	10.0	20.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	6.0	0.0	0.1
10p	0.0	20.0	30.0	5.0	0.0	0.0	20.0	25.0	0.0	0.0	0.0	20.0	25.0	0.0	0.4
1q	0.0	20.0	30.0	5.0	0.0	0.0	20.0	25.0	0.0	0.0	0.0	20.0	25.0	0.0	0.4
3q	10.0	40.0	50.0	25.0	15.0	10.0	40.0	50.0	30.0	20.0	10.0	40.0	40.0	30.0	1.0
99p	21.8	83.6	80.4	50.0	57.2	37.2	91.8	96.8	80.0	60.0	37.2	80.0	80.0	73.6	9.7
max%99	14.7	7.7	24.4	0.0	39.9	168.8	8.9	3.3	25.0	16.7	34.4	18.8	6.2	22.3	174.0
-3sd	-15.2	-15.1	-7.2	-24.2	-31.9	-27.1	-26.0	-23.7	-37.4	-33.5	-22.5	-25.2	-14.9	-36.5	-6.6
+3sd	24.4	78.2	81.1	57.2	51.1	38.7	86.9	94.9	74.6	56.9	35.9	85.5	81.4	73.1	8.6
count3sd	2.0	3.0	2.0	0.0	2.0	2.0	3.0	3.0	4.0	3.0	2.0	1.0	1.0	2.0	2.0

medians were zero and thus Javascript distributions were most positively skewed. Each of the 15 language preference questions contain 1-4 outliers beyond the 3 SD point, but no missing values. Python, R and SAS max values are 12-50% above 99% percentile. The most extreme values (50% or more above the 99th percentile) are found in personal and professional (not industrial) language preference responses; this may be due to the diversity of the student body or possibly influenced by industry surveys.

Table 4: Univariate measures for new course preference degree completion:

	Py take	DE take	AA take	SA take	# taken
count	165.0	165.0	165.0	165.0	165.0
miss	0.0	0.0	0.0	0.0	0.0
min	0.0	0.0	0.0	0.0	1.0
max	100.0	100.0	100.0	100.0	12.0
mean	73.8	57.9	56.1	54.6	6.3
median	82.0	60.0	61.0	54.5	6.0
std	30.3	32.3	33.8	33.0	3.0
skew	-1.1	-0.3	-0.2	-0.1	0.1
kurt	0.1	-1.1	-1.3	-1.3	-0.9
01p	0.0	0.0	0.0	0.0	1.0
05p	7.2	0.0	0.0	0.8	2.0
10p	60.0	30.0	25.0	25.0	4.0
1q	60.0	30.0	25.0	25.0	4.0
3q	100.0	85.0	85.0	81.0	9.0
99p	100.0	100.0	100.0	100.0	12.0
max%99	0.0	0.0	0.0	0.0	0.0
-3sd	-17.0	-39.1	-45.4	-44.3	-2.6
+3sd	164.6	154.9	157.6	153.5	15.3
count3sd	0.0	0.0	0.0	0.0	0.0

The 4 **course preference questions** **in table 4** contain no outliers beyond 3SDs. Students voted a median 82% for the Python course and between 55-61% for the 3 other courses. Course preference distributions are skewed (moderately negative) only for the Python course. **Degree completion responses** indicate the median student respondent is half done: 6 of 12 courses completed. **Bivariate analysis** **(see correlation matrices and scatter plots on the final pages)**

shows strong **positive relationships** (correlations > 65% echo visible relationships in scatter plots) **between context** responses (personal, professional, industrial) for each **language**. There are **negative** correlations **between languages,** but cursory scatter plot review shows this is **mostly noise**, though more sophisticated analysis may show otherwise.

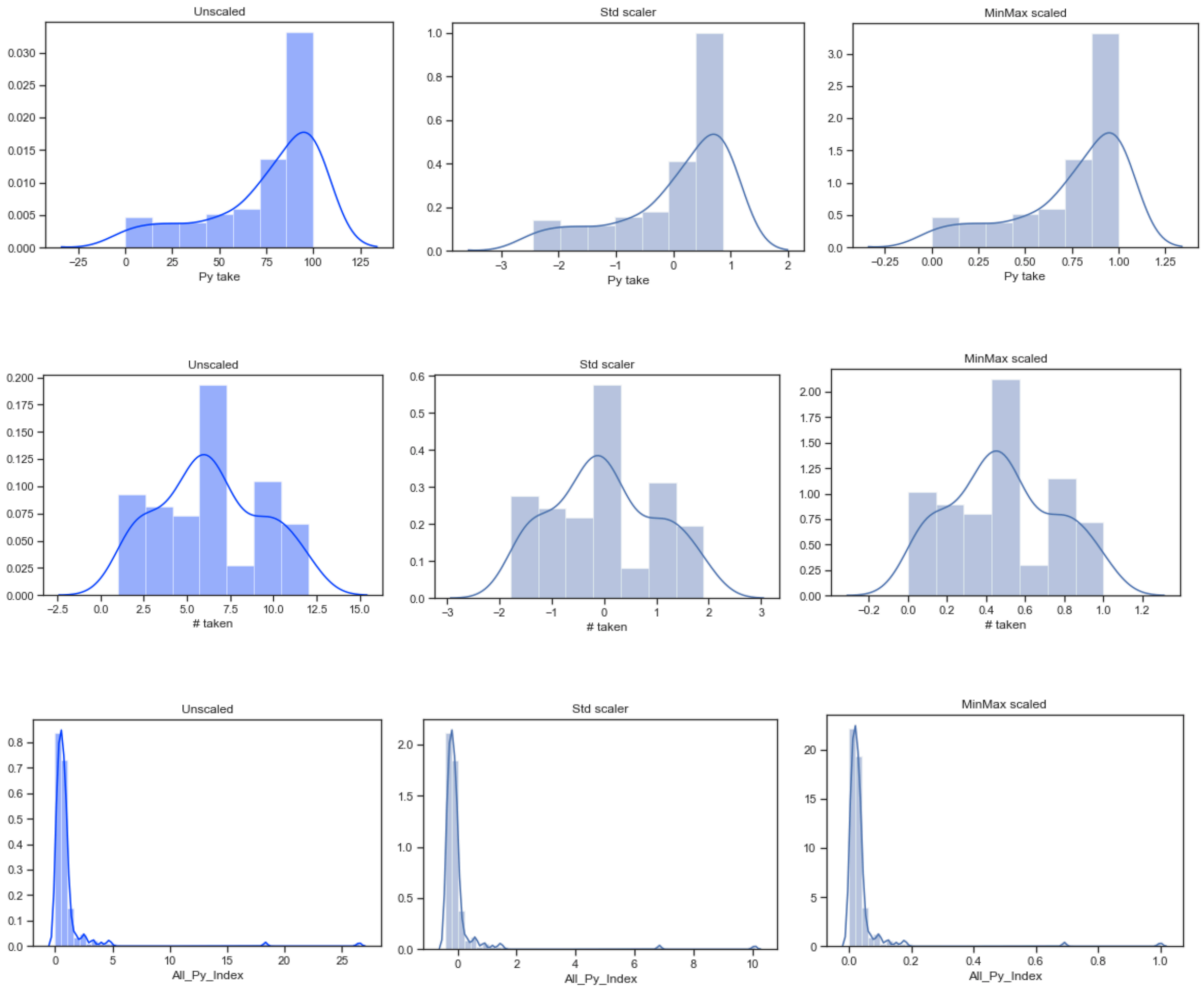
Python codes these steps: (1) **Sklearn's stratified shuffle split** separates the data **80 train/20 test** by employing the courses completed attribute split into 3 bins. This method proved more representative than random separation as shown in **table 5**. (2) **RespondentID** was removed as

Table 5: Stratified vs Random methods
for separating train/test sets

	Overall	Stratified	Random	Rand. %error	Strat. %error
1	0.275362	0.285714	0.285714	3.759398	3.759398
2	0.458937	0.452381	0.500000	8.947368	-1.428571
3	0.265700	0.261905	0.214286	-19.350649	-1.428571

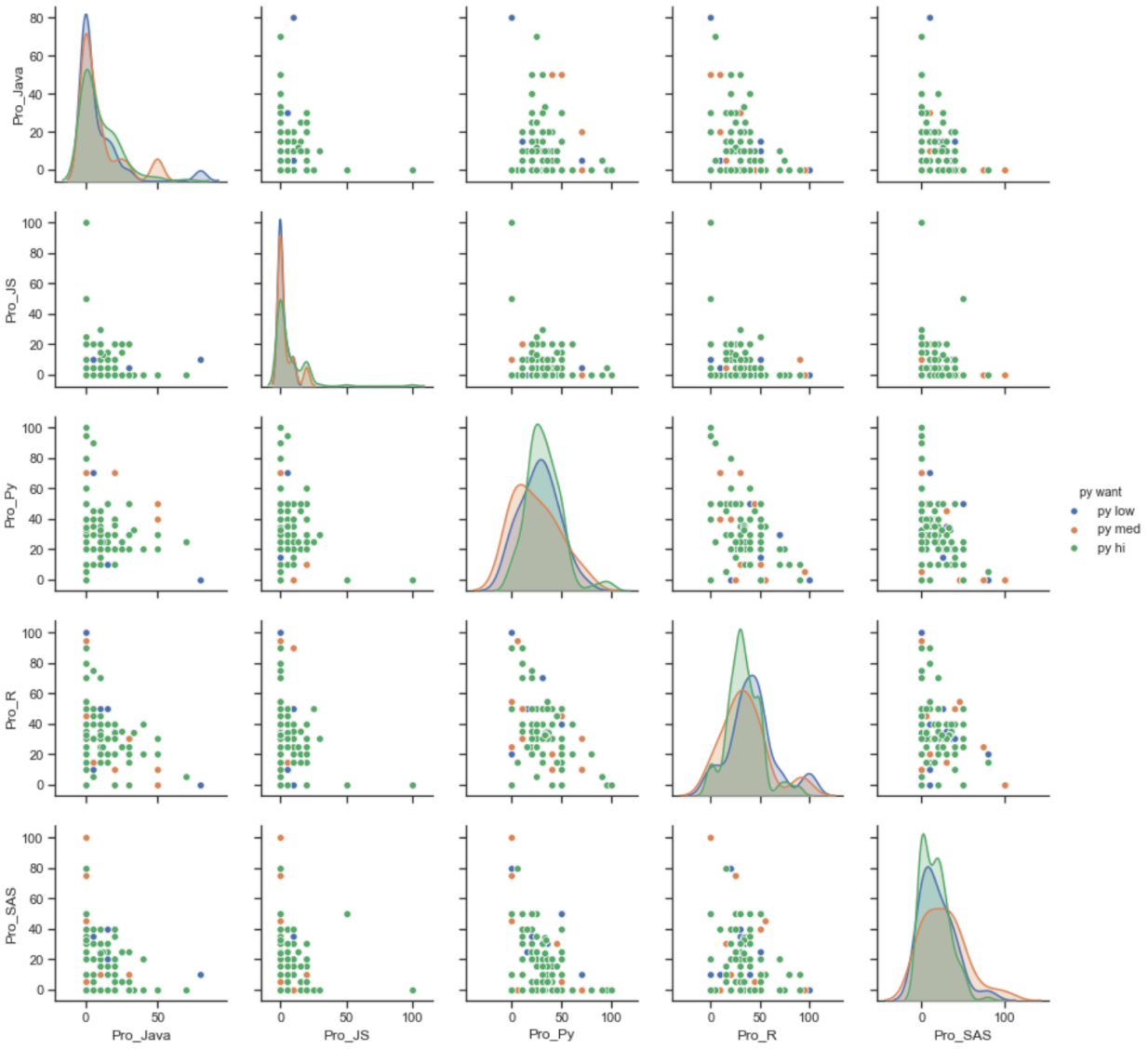
as attribute and used as a unique index identifier; this exposed 2 duplicate rows. Column headings were relabeled to conserve table heading space. (3) **Missing rows** were counted and any attribute missing more than 10% rows discarded. Median and mean are compared in **table 2 above**, with median chosen to **impute** missing values via **sklearn's SimpleImputer. (4)** Via **sklearn's FunctionTransformer**, a **new feature** was **added** = the sum of Python preferences divided into the sum of R and SAS preferences to index the 3 competing data-science languages. (5) 3 **sklearn Pipelines** performed imputation, attribute addition, and 3 forms of scaling: (a) no scaling, (b) standard scaling, (c) minmax scaling. For each of these scalings, **3 attributes are chosen** for their diversity of distribution to see the impact of scaling: **'Py take'** counts students voting for a Python course and is negatively skewed with values ranging zero to 100; **'# taken'** counts number of courses taken in the degree and is fairly normal with no skew

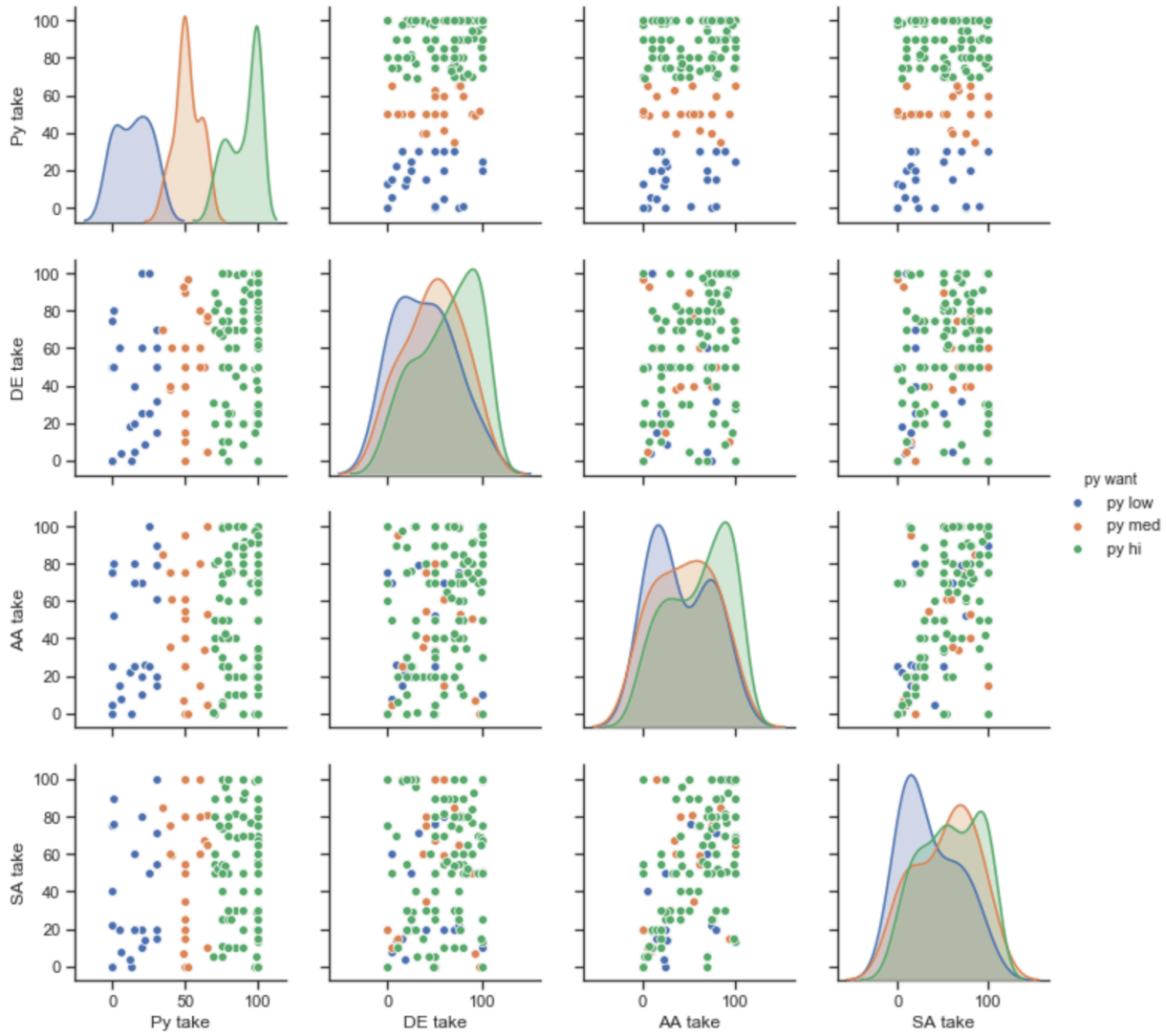
or outliers and ranges 1 to 12; and **'All Py index'** (the new feature) is very positively skewed and ranging mostly zero to 5 with extreme outliers north of 25. In both scaled scaling methods **shown below in the middle and RHS plots**, the shape of the distribution is unchanged, and the scale is made uniform. The 'All Py index' and other skewed distributions with outliers would enjoy normalization or log transformations not shown here.

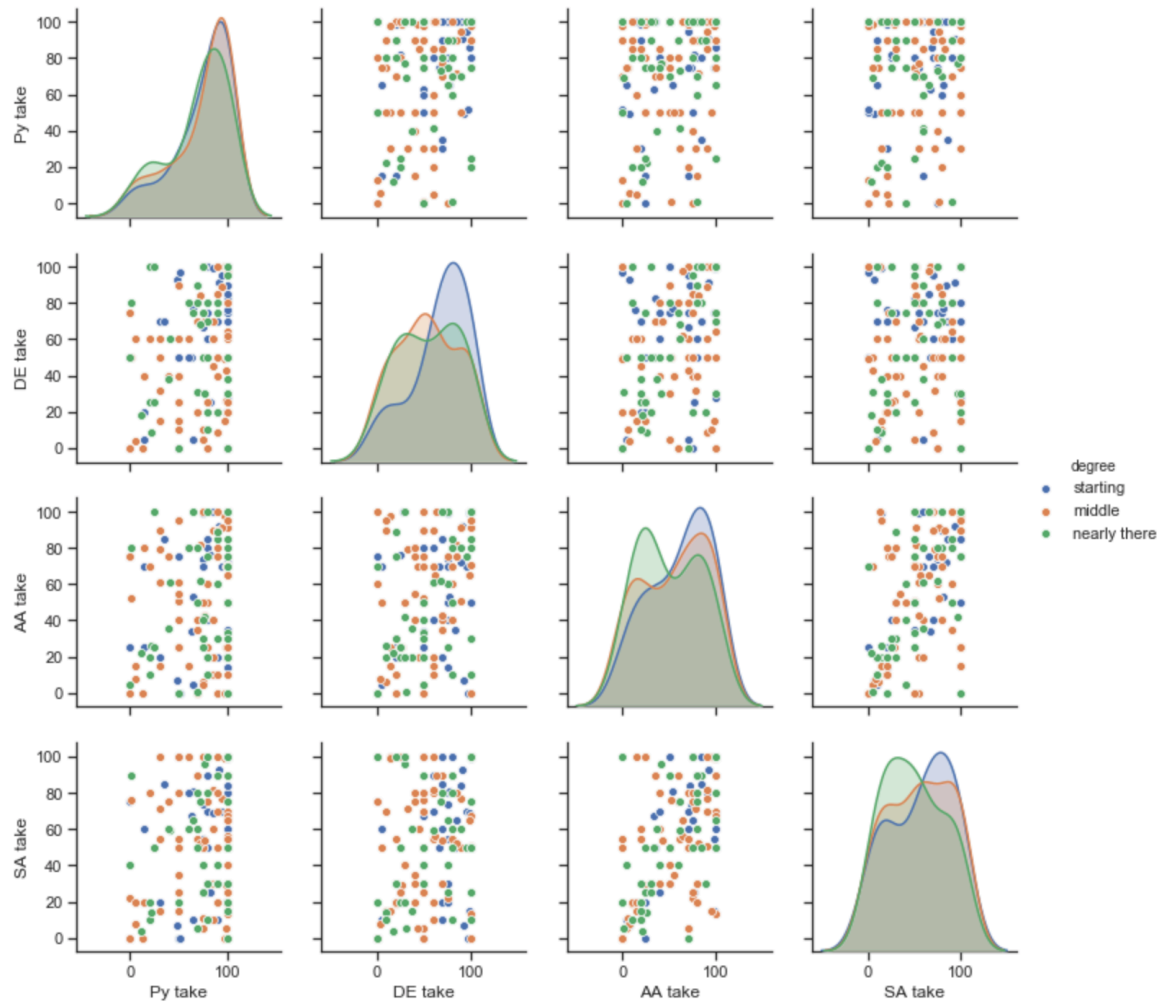


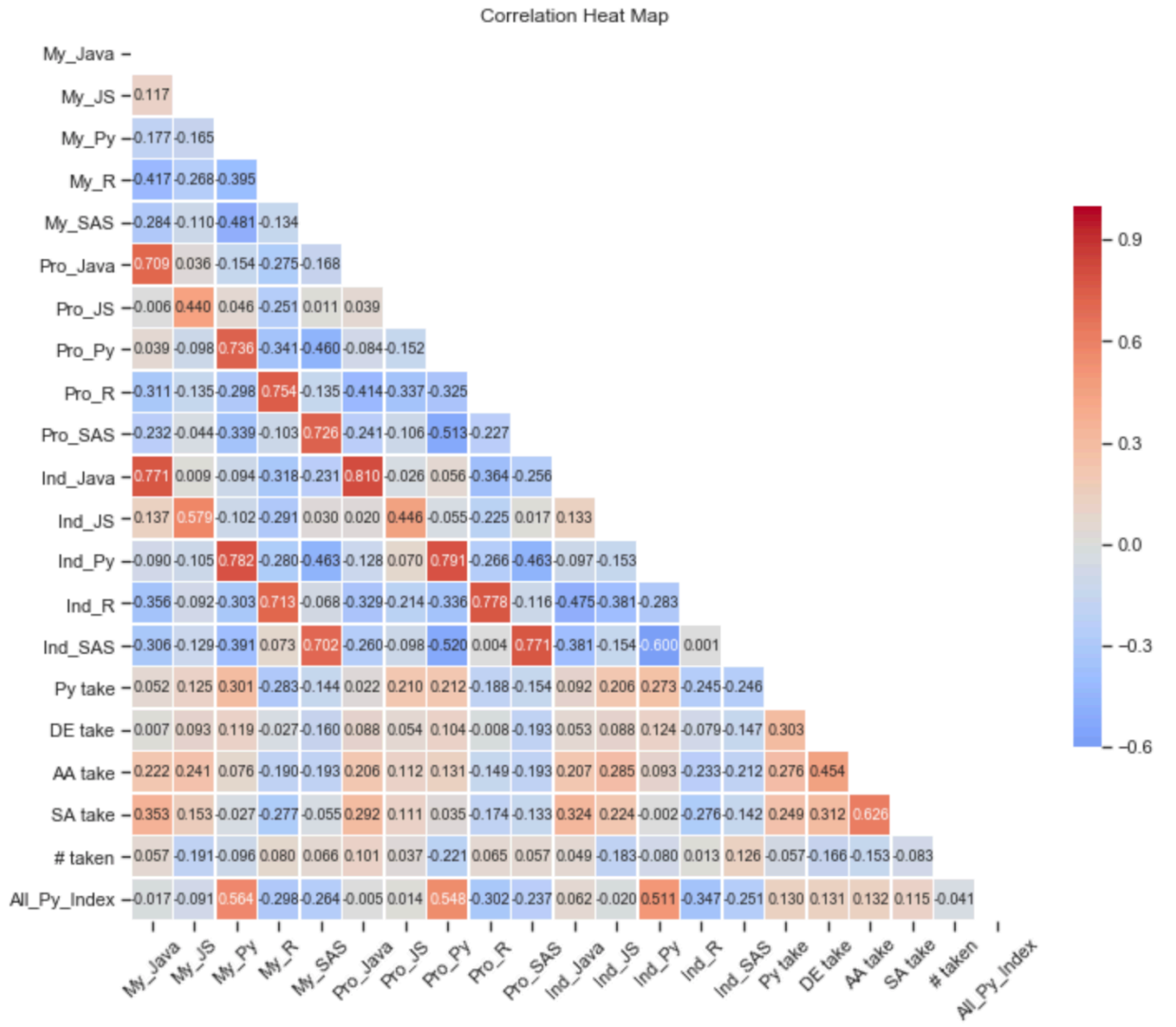
Recommendations for management: Principal conclusions are that (1) the Python course will be very popular and (2) amongst new students, the other new courses or new courses in

general will be popular. (3) **Further analysis should be conducted to investigate relationships amongst responses to these questions.** The 3 sets of scatter plots and distributions along the diagonal support these 3 points. (1) The first plot below shows professional interest in Python and R versus SAS and Java and relatively little linear relationship across views on these 5 language groups as mentioned in the univariate analysis above. These plots of professional preferences are similar to industrial and personal preferences, and all indicate a **strong interest to learn Python and R**. (2) the second plot shows interest in the 4 course offerings, binned by strength of interest in the Python course. The first column is unimportant but the other 3 show modal differences depending on interest in the Python course: those interested in the Python course are also most interested in the DE, AA and SA course offerings. This may demonstrate students' interest in new courses in general rather than any particular course interest, or it may show Python is correlated with new course offering interests. Binning the other course offerings or other more sophisticated analysis may indicate an answer. (3) the final set of plots shows a strong interest in Python in the first column, and shows that newly started degrees are more interested in new course offerings than students nearly done with the program. This shouldn't surprise but may indicate the courses will be popular amongst new or prospective students.


Pro_Java
Pro_JS
Pro_Py
Pro_R
Pro_SAS
py want
py low
py med
py hi


Py take
DE take
AA take
SA take
0
20
40
60
80
100
0
50
100
py want
py low
py med
py hi


Py take
DE take
AA take
SA take
0
20
40
60
80
100
degree
starting
middle
nearly there


Correlation Heat Map
My_Java
My_JS
My_Py
My_R
My_SAS
Pro_Java
Pro_JS
Pro_Py
Pro_R
Pro_SAS
Ind_Java
Ind_JS
Ind_Py
Ind_R
Ind_SAS
Py take
DE take
AA take
SA take
taken
All_Py_Index
0.9
0.6
0.3
0.0
−0.3
−0.6